含意推論タスクにおいて大規模言語モデル(LLM)の出力は対話相手の特性の影響を受けるか?

Are outputs of the Large Language Model affected by addressees' characters in the task of scalar implicature?

西畑 千哲[†], 小林 春美[‡], 安田 哲也[§]
Chisato Nishihata, Harumi Kobayashi, Tetsuya Yasuda
[†]東京電機大学大学院, [‡]東京電機大学, [§]東京大学
Graduate School of Tokyo Denki University, Tokyo Denki University, The University of Tokyo
23rmd37@ms.dendai.ac.jp, h-koba@mail.dendai.ac.jp, t-yasuda@g.ecc.u-tokyo.ac.jp

概要

本研究では，大規模言語モデル(LLM) がどのような含意推論を出力するかについて調べた．利用した LLM は OpenAI 社の chat GPT-3.5 であった．この GPT-3.5 に実験参加者としての役割を与え，Nishihata, Kobayashi, & Yasuda (2023)の手順を利用し，含意推定に関するタスクを行った．人間の実験参加者のデータ(Nishihata et al.)と比較した結果，GPT-3.5 では人間とは異なり，コミュニケーション相手によって含意推論を変えない場合が多い可能性と，文脈が交絡した場合，文脈情報を利用しない可能性が示された．

キーワード : 大規模言語モデル(LLM)，chatGPT-3.5，含意推定，ロボットエージェント

1. 目的

現在，Large Language Model (LLM: 大規模言語)が台頭し，ロボットに実装し，人との相互作用を調べようとする試みがある(Kim et al., 2024)．また，LLM の語用論における推論能力を明らかにしようとする研究 (Bojic et al., 2023)や Grice の協調原則のフレームワークから調べる研究 (Park et al., 2024)がある．本研究では，chatGPT に実験参加者の役割を人間として与え，エージェント(i.e., robot)や人間の発話 (i.e., プロンプト)に対し，含意をどの程度推定するのかを調べ，人間のデータ(Nishihata et al., HAI2023)と比較した．

Grice の協調原則に基づく 4 つの格率 (Grice, 1989)は，語用的推論の処理を考えるうえで重要な指標になり得る．4 つの格率は量，質，関係，様態があり，会話の相手に与える情報について，過不足のない情報であること，偽りなく正確な情報であること，関連した情報であること，これらの情報を適切な形式で伝えることを要請する．これらの格率は含意を導出する際に寄与する(Horn, 2013)．含意とは文のそのものの意味だけでなく，他の意味を包含することである．たとえば，「It's hot today.」という発話は字義通りの解釈を促すのに対し，「It's warm today.」という発話では「hot]という語を選ばず「warm」という語を選んでいる事実から，「hot というほどではない」という含意が含まれることになる．これは尺度含意(scalar implicature)と呼ばれている(Horn, 2013)．尺度含意 (量的含意と言われる場合もある)は，弱い意味(e.g., warm, some)を持つ語と強い意味 (e.g., hot, all)を持つ語で構成されている．

Nishihata et al. (2023)は，量的含意を利用し，人間が発話する場合とロボットエージェントが発話する場合に，参加者である人間がどのようにそれぞれの発話において量的含意を推測するのかを調べた．なお，Nishihata et al.は他者の内的状態の推定程度が含意解釈に寄与するという予測のもとに，ロボットエージェントの内的状態を推定しやすい場合(アルゴリズムがわかりやすい; 100%のキー操作反映) / しにくい場合(50%のキー操作反映)の違いを調べた．その結果，求められた量は文脈の影響を受けた量を推定していた．エージェントに着目してみると，内的状態を推定しにくいロボットと人間エージェントを比較した場合は，参加者はほぼ同じ量を見積もった．しかしながら，内的状態を推定しやすいロボットの方が内的状態を推定しにくいロボットを比較した場合，参加者は，量の見積もりはロボットの違いに影響された．これは予測を支持し，人間は含意を導出する際に，会話相手が持つ特徴の影響を受けていた可能性を示唆する．

現在，LLM が人間の言語能力と同等になり得るのかを多くの研究者が調べている．Bojic et al. (2023)では，文脈や暗示的意味を考慮する必要がある表現を解釈する能力を調べるために，複数の LLM を利用し，Grice の格率に関する質問をした．タスクに関しては，Grice の格率や関連性理論 (Sperber & Wilson, 1995)に基づくものであり，正

誤を判断させた．例えば，A young man approaches a person at a party and says “Hello, my name is Stefano”. The person replies: “Mine isn’t”というプロンプトは. The person replies: “Mine isn’t”という箇所が, 正しい名前の情報に言及していないため量の基準に違反する．結果として，LLM，特にchatGPT 4.0 は適切に解釈を行い，反応速度も速いことが示された．これらの評価は，Grice や Sperber and Wilson の理論に基づいて LLM を評価した結果であると考えられる．Nishihata et al. (2023)でも Grice の枠組みから人間を参加者として研究を行っているが，LLM を「参加者」として回答させた場合，Bojic et al. の知見と同様の知見が得られるのだろうか．LLM にキャラクターを演じさせる研究には，RPG に見立てたもの(Park et al., 2023)や，物語のシナリオをプロンプトに使用したもの(Xu et al., 2024)などがあるため, LLM が参加者の役割を演じれる可能性はある．

本研究では chatGPT を使用し，特に尺度含意について LLM ではどのような推論が行われるのかを調べた．人間と同じような解釈, 出力が行われるのか, それとも人間とは全く違う解釈, 出力が行われるのか, すでに得られている人間のデータと比較した．実験手順は，Nishihata et al. (2023)をもとにプロンプトを設定した．実験では chatGPT は参加者の役割を与えられた上で，未知の乗り物にエネルギーがすでに入っているが，その量はわからないという場面を設定され，その乗り物に入れるべきエネルギー量を決定する課題を行った．参加者は乗り物に入っているエネルギーの量について「十分ある」または「ほとんどない」ということばで伝えられ, この乗り物にエネルギーを「たくさん」または「少し」入れて欲しいと登場したエージェントにより頼まれた．その後補給するエネルギーの量を，数値として読み上げるよう指示された．

人間の場合の結果では, 少ない量を要求された場合，人間に要求された時とロボットに要求された時では，補給するエネルギー量に差があった．では GPT-3.5 では，コミュニケーションの相手によって，含意の解釈は変わるのだろうか．また，LLM は学習データから，文脈を考慮した処理を行うはずだと考えられるが, エネルギーが「十分ある」のに「たくさん」要求されるような，言語と状況が不一致しているように見える（コンフリクト）場合に，人間と同じような値を出力するのだろうか．chatGPT では，特性を指定するとその役になりきったような挙動を見せることが報告されている. そのため, コミュニケーションの相手によって，推論を使い分けてエネルギー量を調整するだろうと予測した. また, 状況がコンフリクトした場合も，文脈を考慮し人間データと同じようなエネルギー量を出力すると予測した.

2. 方法

2.1. 使用したデータセット

LLM データ (16 chats, chatGPT-3.5), 人間データ (13 人, Nishihata et al., HAI2023)であった．LLM は chatGPT-3.5 (OpenAI; 2024 年 4 月使用)を利用した.

2.2. プロンプト

HAI のデータは，13 人の理系の大学生，大学院生であった．参加者は，PC に表示された人，ロボット，ロボットアームの３種類のキャラクターを操作し，その動き方を経験した. その後, 画面に人間またはロボットが表示され，「(人間/ロボット）は次のように言っています．エネルギーの量を決めてください.」と実験者に指示された. セリフの例は，文脈条件 Abundant と指示文条件 Little の場合では，「この乗り物にはエネルギーが十分あります」「エネルギーを少し入れてください」と表示された. 参加者は未知の乗り物に補給するエネルギー量を PC を操作し決定した．本実験では、ChatGPT に参加者になりきらせるためのプロンプトを設定した．また，HAI の手順に沿うように，登場した人間とロボットをモデレーターとして設定し，その特性を記述した.

図 1 使用したプロンプト

```
#命令文
あなたは，#実験参加者になりきって，#制約条件に従って，次の回答から，私の質問に答えてください.
私が#ロボットモデレーターと#人モデレーターを演じます.
#制約条件
*空欄の箇所がある場合は，適切に設定を作り出して回答すること
*キャラクターの特性を分析して対象の視点から回答すること
*対象の特性に基づいた情報や視点から回答すること
*あなたは，実験参加者で，エネルギーを補給する人です.
*エネルギーは，液体ですが，未知なエネルギーです.
*あなたは，エネルギーを補給できる場所にいます.
*エネルギーは，無限です.
*あなたは乗り物にどの程度エネルギーが入っているかを知ることができません.
*あなたは0から３００まで補給するエネルギーを選び，数値を伝える義務があります.
#実験参加者の設定
*名前：{}
*年齢：{20}
*性別：{男性}
*種族：{人間}
*役割：{エネルギーを入れる担当}
*性格の特徴：{}
*言葉遣い：{}
*国籍：{日本}
```

2.3. 手順

ChatGPT に制約条件，実験参加者の特性，モデレーターの特性を送信した. 次に, 実験参加者を演じているかを確認した．その後，(ロボット A:「この乗り物にはエネルギーが十分あります.」ロボット A:「エネルギーを少し入れてください.」あなたは，補給したエネルギー量を読み上げてください．) と送信し，出力された値を記録した.

2.4. 条件

条件は，データ条件 (Data: Human (HAI)/LMM (chatGPT))，モデレータ条件 (Agent: Human/Robot)，文脈条件(Context: Abundant (十分あります) / Scarce (ほとんどありません))，指示文条件(Requesting: Much (たくさん) / Little (少し))の 2 つであった．実験参加者として役割を与えられた GPT は，年齢が 20 歳で統一され，男性または女性の性別が与えられた．

2.5. 分析方法

実験参加者として役割を与えられた chatGPT に，補給したエネルギー量を読み上げさせ記録した．統計モデリングには，R software を用いた．

参加者がエージェントの主体性や協力的な体験の存在が含意推論に関わっているかどうかを調べるために，統計モデルを使ってエネルギー量を推定した．統計モデルには，線形混合モデル（LMM: Liner Mixed Model）を用いた．この統計モデルは，lme4 パッケージの lmer 関数を用いて構築した．まず，実験条件とその交互作用を固定効果，個人差と人間の主体性の影響をランダム効果として含む最大モデルを構築した．次に，赤池情報量規準に基づく前向きのステップワイズ法を用いて，データに適合するモデル候補を検討した．

結果，分析に用いるモデルは固定効果として，Data，Agent，Context，Requesting およびそれらの相互作用を適用することが提案された（lmer(The-amount-of-energy ~ Data × Agent × Context × Requesting + (1|Participants) + (1|Trial))）．なお，統計モデルで交互作用効果が認められた場合，ダミーコーディングによるポストホック検定で単純主効果を検討した．

3. 結果

図 2 に各条件におけるエネルギーの量を示す．各エネルギーの量を従属変数とした LMM の結果を表 1 に示す．Intercept ($\beta = 153.10, t_{27.00} = 44.35, p < .001$)，Data ($\beta = -20.21, t_{27.00} = -2.93, p < .001$)，Context ($\beta = -39.40, t_{653.00} = -14.22, p < .001$)，Requesting ($\beta = 179.85, t_{653.00} = 64.89, p < .001$)，Data と Context 間の相互作用 ($\beta = -52.65, t_{653.00} = -9.50, p < .001$)，Data と Requesting の相互作用 ($\beta = -72.48, t_{653.00} = -13.08, p < .001$)，Context と Requesting の交互作用 ($\beta = -72.48, t_{653.00} = -13.08, p < .001$)，Data，Agent，Requesting の二次の交互作用 ($\beta = -23.58, t_{653.00} = -2.13, p = .034$)，Data，Context，Requesting の二次の交互作用 ($\beta = -44{,}34, t_{653.00} = -4.00, p < .001$)は有意であった．しかし，Agent ($\beta = -0.97, t_{653.00} = -0.35, p = .73$)は有意でなかった．

Human/Little では，Human に指示された方が，LLM に指示された時より，有意に多くエネルギー見積もった ($t_{653.00} = 2.08, p = .039$)．しかし，LLM/Little では，Human に指示された方よりも，Robot に指示された時の方が，エネルギー量を多く見積もる傾向であった($t_{653.00} = -1.69, p = .091$)．また，Abundant/Much の時，LLM の方が，Human よりも有意に多くエネルギーを見積もった($t_{58.86} = -11.16, p < .001$)．さらに，Scarce/Much の時，LLM の方が，Human よりも有意に多くエネルギーを見積もった($t_{58.86} = -2.27, p = .027$)．Scarce/Little の時，LLM の方が，Human よりも有意に少なくエネルギーを見積もった($t_{58.86} = 3.72, p < .001$)．

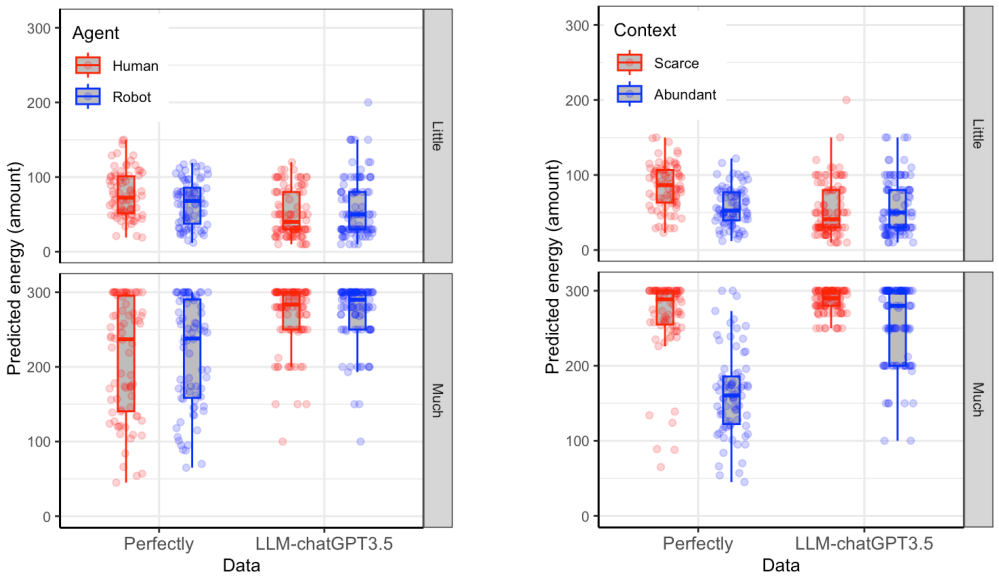
4. 考察

本研究は，GPT-3.5 に人間と同様の含意推論を行うかに関して，Nishihata et al. (2023)のデータと比較した．

表 1: 線形混合モデルを用いた結果

Fixed effects:						
	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	153.096	3.4517	27	44.354	< 2e-16	***
Data	-20.2143	6.9033	27	-2.928	0.00685	**
Agent	-0.9667	2.7717	653	-0.349	0.72736	
Context	-39.4002	2.7717	653	-14.215	< 2e-16	***
Request	179.853	2.7717	653	64.89	< 2e-16	***
Data:Agent	9.1707	5.5434	653	1.654	0.09854	.
Data:Context	-52.6546	5.5434	653	-9.499	< 2e-16	***
Agent:Context	-7.5954	5.5434	653	-1.37	0.1711	
Data:Request	-72.4816	5.5434	653	-13.075	< 2e-16	***
Agent:Request	-5.1418	5.5434	653	-0.928	0.35398	
Context:Request	-54.5264	5.5434	653	-9.836	< 2e-16	***
Data:Agent:Context	-9.5657	11.0867	653	-0.863	0.38856	
Data:Agent:Request	-23.5753	11.0867	653	-2.126	0.03384	*
Data:Context:Request	-44.3446	11.0867	653	-4	7.06E-05	***
Agent:Context:Request	-6.9663	11.0867	653	-0.628	0.52999	
Data:Agent:Context:Request	-3.016	22.1734	653	-0.136	0.89185	

図 2 見積もったエネルギー量



GPT-3.5 は文脈が利用できる「ほとんどない」と指示文「たくさん」の場合に，人間よりも少なく量を見積もった．このことを考えると，GPT-3.5 は文脈を考慮していないことが推察される．しかしながら，GPT-3.5 は「十分ある」という文脈があり，指示文が「少し」の場合に人間と同様に量の推定を行っていた．この文脈においては，文脈情報の指示文への寄与はあるため，なぜ GPT-3.5 が人間と同様のパフォーマンスを示していたかを調べる必要がある．

GPT-3.5 は指示文が「少し」の時に，ロボットエージェントの方が人間エージェントよりも多く量を見積もっていた傾向にあった．この文脈において，人間は人間エージェントの方がロボットエージェントよりも多く量を見積もっていた．可能性として，自身の属性の方が量の推定を見積もりやすいため，人間なら人間エージェントの方が量を見積もりやすいことが考えられる．LLM は，言語データによる学習がなされているため，人間を模していると考えることも可能である．しかしながら，人間は状況に合わせて意図的に言語を使用することや，視線，表情，ジェスチャーなど非言語情報による意図明示とともに提示する場合がある．このような意図明示情報は，LLM で学習できるデータには明示的には含まれない．よって，この点を明らかにする必要は大いにある．

含意をどの程度推定するのか，人間の実験参加者のデータと比較した結果，GPT-3.5 では文脈が交絡した場合，文脈情報を人間の程度に比べて適用しない可能性があった．今後，人は文脈情報を過剰に利用しているのか，GPT-3.5 が過度に利用していないのかに関して，検討する必要がある．また，GPT-3.5 以外の LLM での検証や，今回使用したプロンプトは Zero-shot prompting に近い，自由回答の形をとっていた．プロンプトによる学習方法や回答方法を検討することも重要である．さらに，LLM の身体性について，Hardy et al. (2023)は CogSci 2023 でのワークショップにて，LLM は身体性をもつ人間の言語を学習データとしているため，完全に身体性がないとは言い切れないという議論を行っていた．また，実際に LLM を身体性を持つヒューマノイドに実装させた試みもある(Yoshida et al., 2024)．この身体性の議論は，今後も行う必要がある．

謝辞

本研究は科研費 JP20H01763 (H.K.)，科研費 JP20K03375 (T.Y.)の助成を受けた．

主要参考文献

Bojic, L., Kovacevic, P., & Cabarkapa, M. (2023). GPT-4 Surpassing Human Performance in Linguistic Pragmatics. *arXiv preprint arXiv:2312.09545*.

Grice, P. (1989). *Studies in the Way of Words*. Harvard University Press.

Kim, C. Y., Lee, C. P., & Mutlu, B. (2024, March). Understanding Large-Language Model (LLM)-powered Human-Robot Interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 371-380).

Nishihata, C., Kobayashi, H., & Yasuda, T. (2023, December). Human-like “agents” or “tools”?: Exploring the implicature-of-quantity in HAI. In *Proceedings of the 11th International Conference on Human-Agent Interaction* (pp. 387-389).

Park, D., Lee, J., Jeong, H., Park, S., & Lee, S. (2024). Pragmatic Competence Evaluation of Large Language Models for Korean. *arXiv preprint arXiv:2403.12675*.

Yoshida, T., Baba, S., Masumori, A., & Ikegami, T. (2024). Minimal Self in Humanoid Robot" Alter3" Driven by Large Language Model. *arXiv preprint arXiv:2406.11420*.